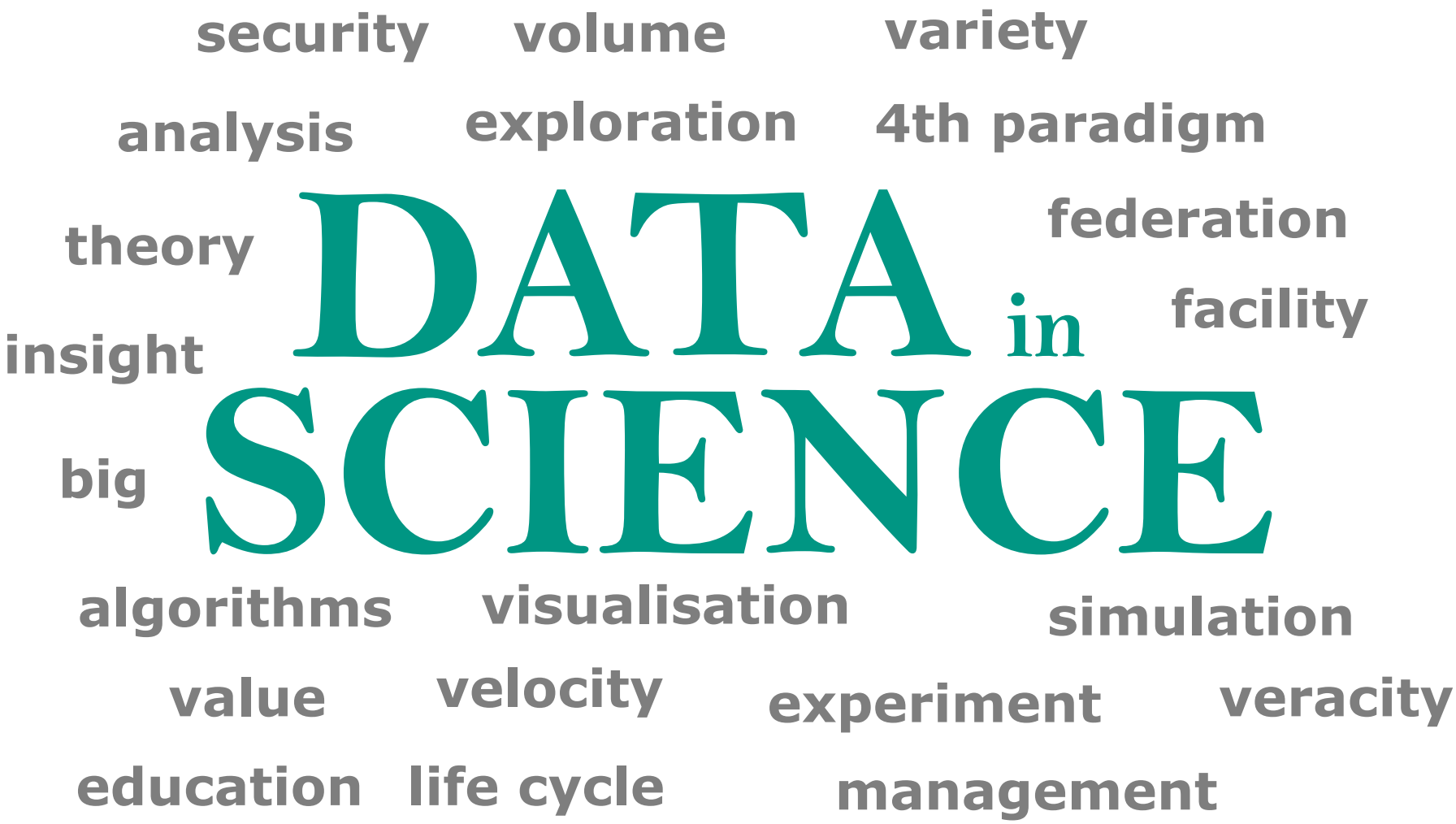
Karlsruher Institut für Technologie

Unterstützung datenintensiver Forschung am KIT – Aktivitäten, Dienste und Erfahrungen

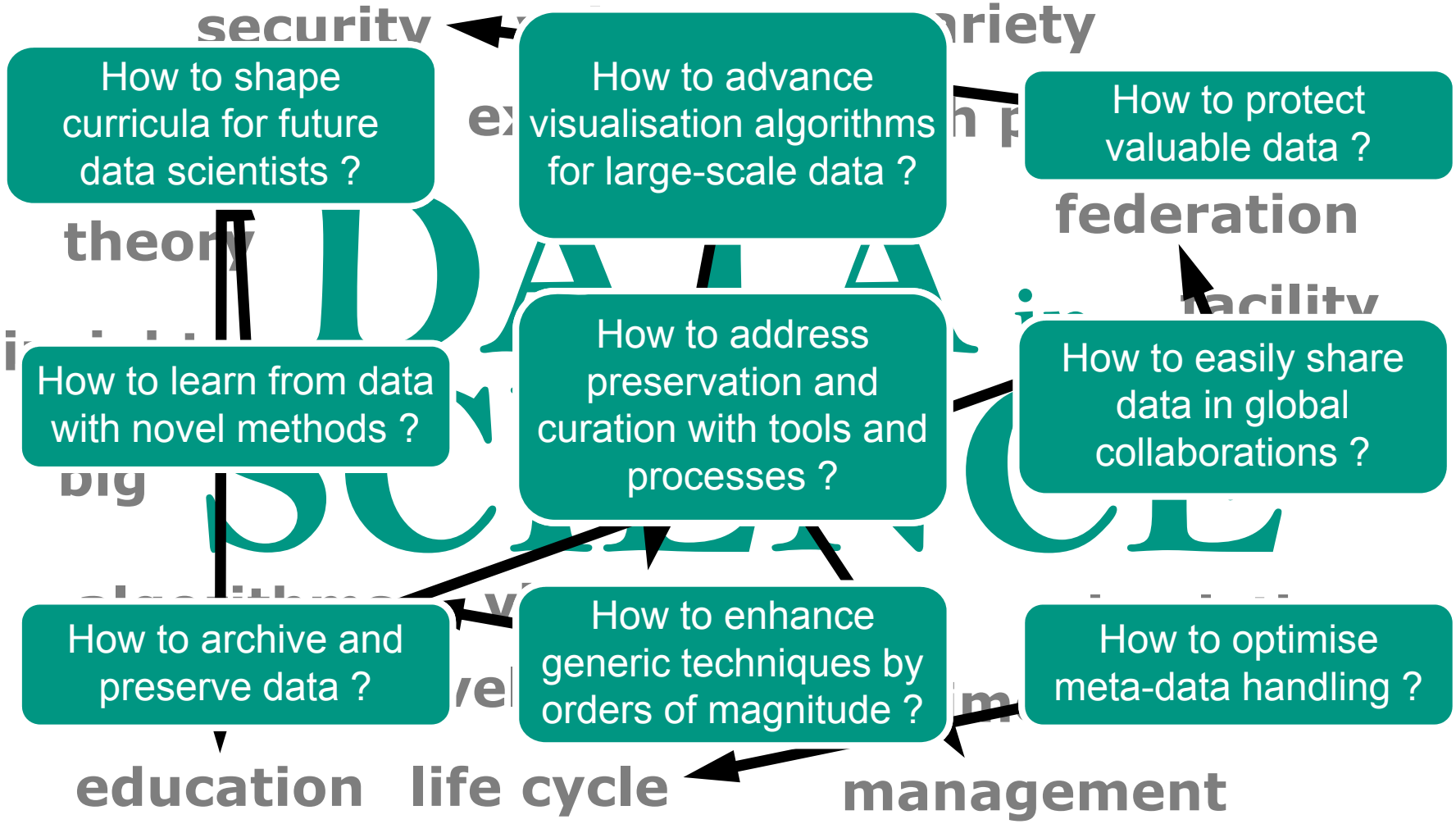
Achim Streit

Steinbuch Centre for Computing (SCC)

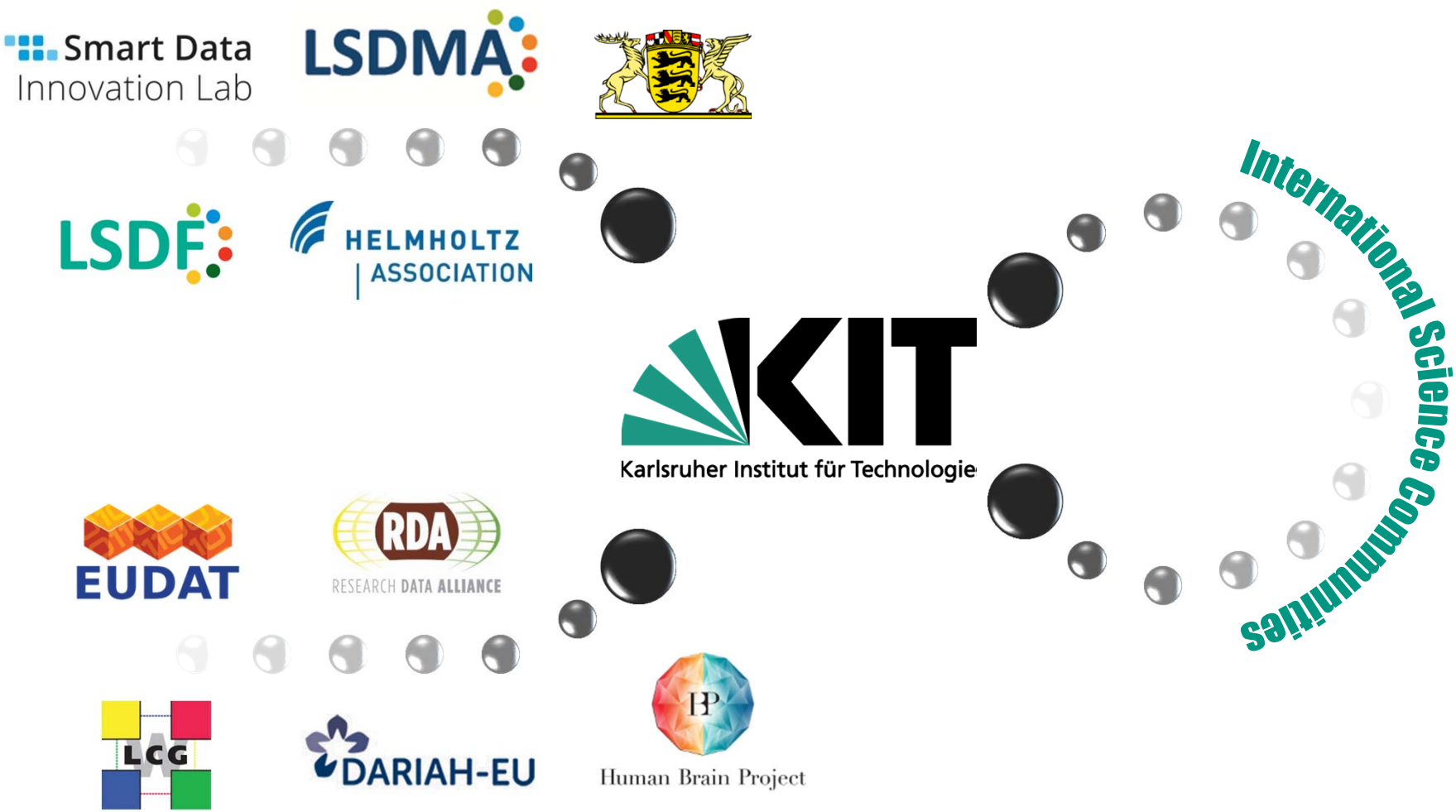
The Challenge of Big Data in Science



The Challenge of Big Data in Science

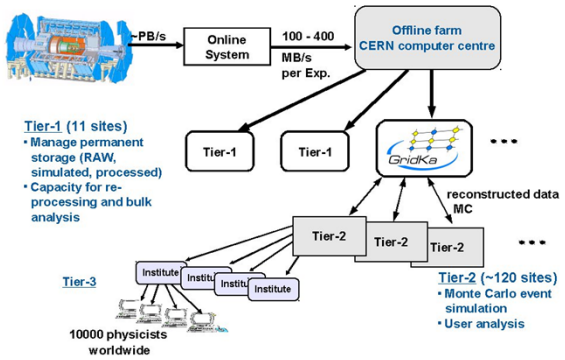


Collaboration is key



GridKa

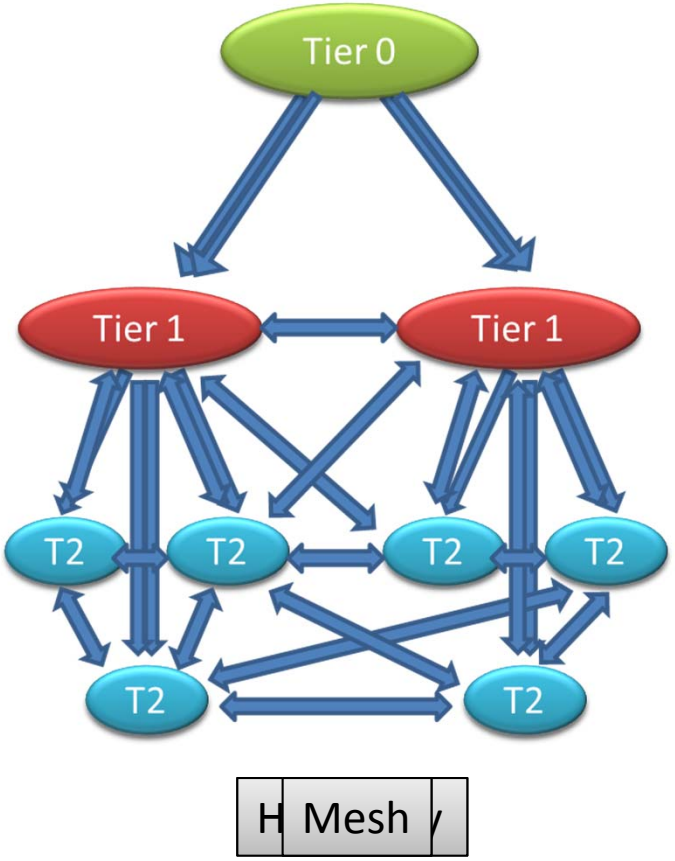
- German Tier-1 Center in WLCG
  - Supports all LHC experiments + Belle II + several small {HE/AP}P communities
  - Benchmarked reliability of 99.5%
- Resources
  - >10,000 cores, utilization >95%
  - Disk space: 12 PB, tape space: 17 PB
  - 6x10 Gbit/s network connectivity
- 14% of LHC data permanently stored at GridKa
- Serves > 20 T2 centres in 6 countries
- Services
  - File transfer
  - Regional workload management, file catalog
- Annual international GridKa School
- Global Grid User Support (GGUS) for WLCG



GridKa School

GridKa Experiences

- Evolving demands and usage patterns
  - No common workflows
- Hardware commodity, software not
- Hierarchical storage with tape challenging
- On-site experiment representation highly useful
- Adoption of grid computing outside of HEP rare
- Data access a central issue
  - Random data access by compute jobs
  - Reprocessing

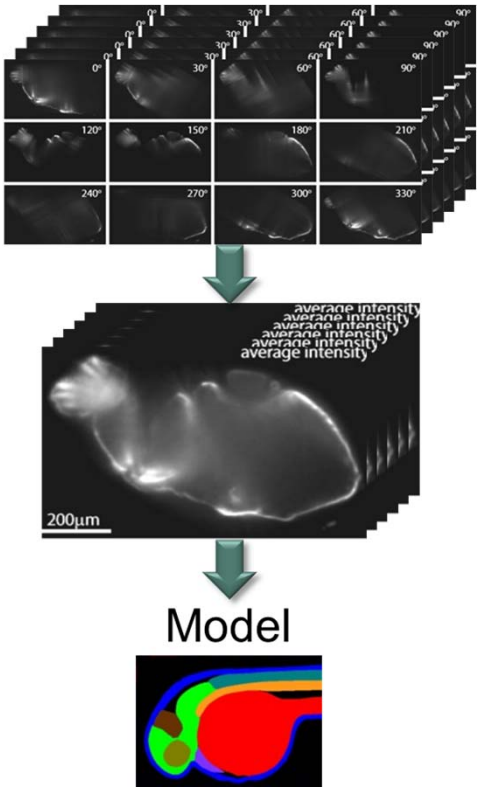


Hierarchical model relaxed

Courtesy of Ian Bird, CERN

Large Scale Data Facility

- Main goals
  - Provision of storage for multiple research groups
    - Systemsbiology, climatology, synchrotron research/materials science, humanities, geo/earth science, …
  - Basis for BW-wide data services
- Resources and access
  - 6 PB of on-line storage (ext. to > 10 PB)
  - 6 PB of archival storage
  - 100 GbE connection between LSDF@KIT and U-Heidelberg
  - Hadoop analysis cluster of 58*8 cores
  - Connection to HPC clusters
  - Jointly funded by Helmholtz Association and state of Baden-Württemberg



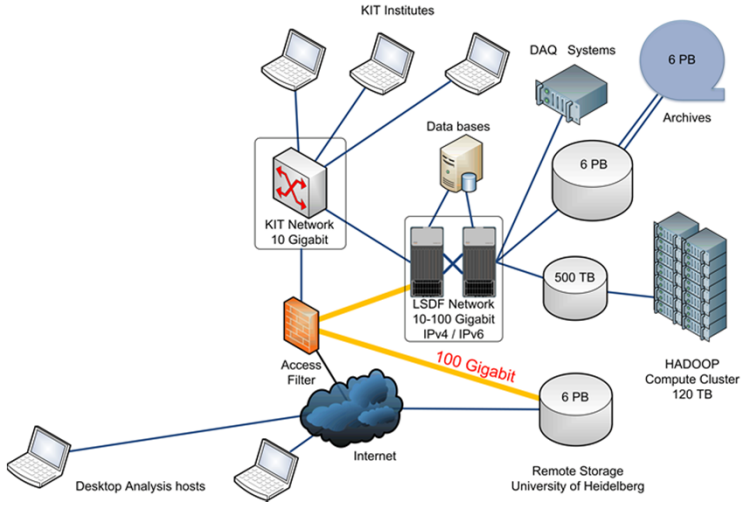
LSDF Experiences

- High demand for storage, analysis and archival
- Research groups vary in
  - Research topics (from genetic sequencing to geophysics)
  - Size
  - IT expertise
  - Need for services and protocols
- Important needs common to many user groups
  - Sharing data with other groups
  - Data safety and preservation
  - ‘Consulting’
- Many small groups depend on LSDF

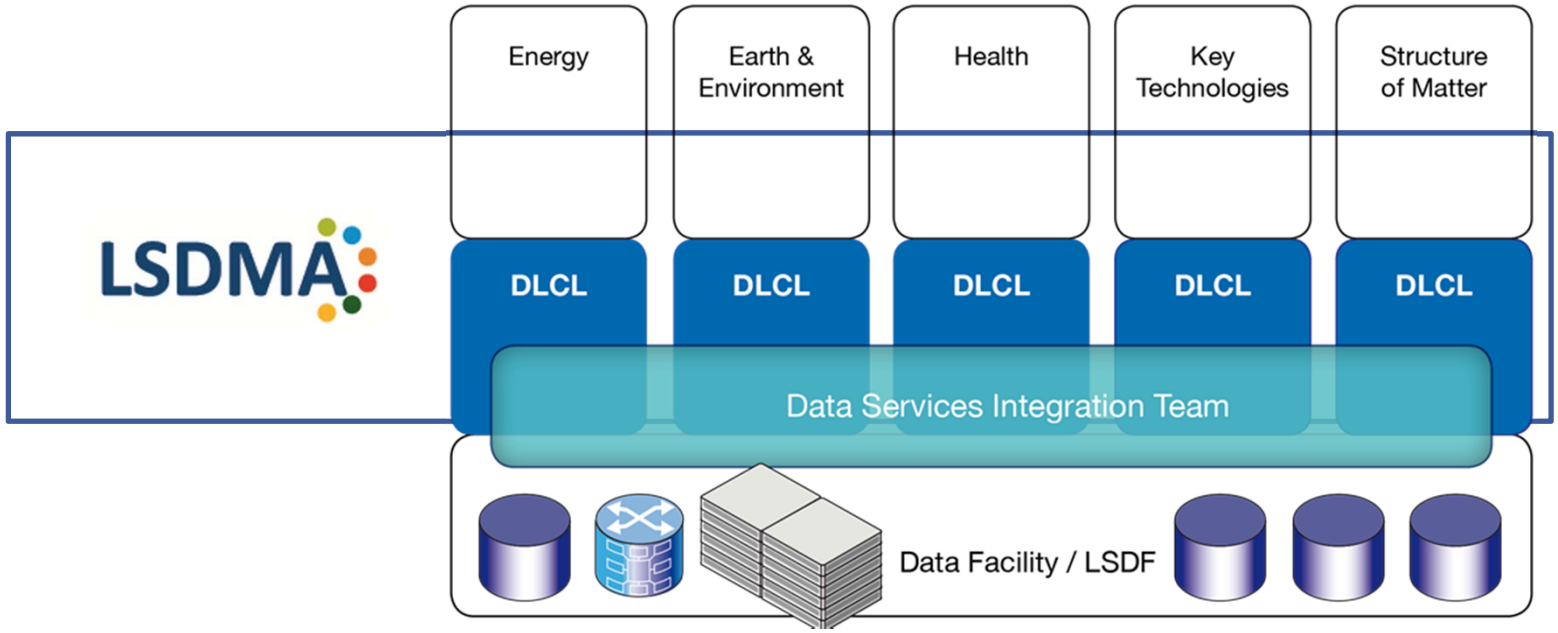
bwLSDF – Services for users in the state of Baden-Württemberg

- bwSync&Share
  - For employees and students at Universities in BaWü
  - Dropbox-like data storage
  - KIT cloud – on-premise solution
  - Access via webbrowser, clients for windows, linux and mobile phones
  - Web based SAML authentification (Shibboleth) - bwIDM
  - Data can be shared with users not registered for the service
  - Service started on 1.1.2014
  - Available to 450.000 users

- bwFileStorage
  - Via SFTP/SCP
  - SAML authentication



Large-Scale Data Management and Analysis – Dual Approach



Data Life Cycle Labs

Joint R&D with scientific user communities

- Optimization of the data life cycle
- Community-specific data analysis tools and services

Data Services Integration Team

Generic methods R&D

- Data analysis tools and services common to several DLCLs
- Interface between federated data infrastructures and DLCLs/communities

Facts and Figures

- Helmholtz portfolio extension
- Initial project duration: 2012-2016
- Partners:

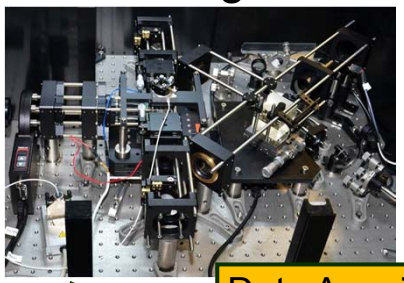
- Sustainability
  - Inclusion of activities into Helmholtz programme-oriented funding (PoF) “Supercomputing & Big Data” from 2015 onwards
  - Cross-programme iniative with other Helmholtz research fields
- Annual international symposium



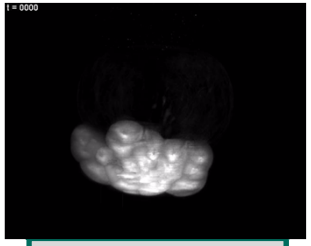
Selected Scientific Highlights

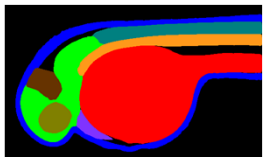
- **DLCL Key Technologies** (KIT, U-Heidelberg, U-Dresden)
  - Light Sheet Microscopy

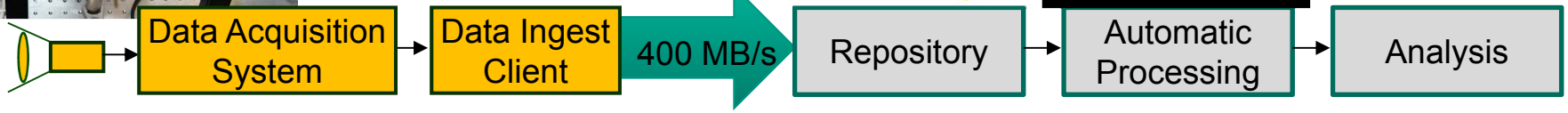
Mikut et al. (2013), Automated processing of zebrafish imaging data - a survey, **Zebrafish 10(3)**, DOI:10.1089/zeb.2013.0886



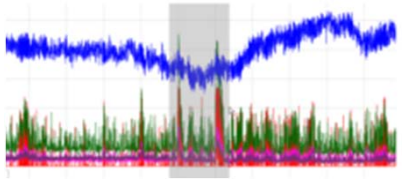
Experimental Microscope 16 TB/d

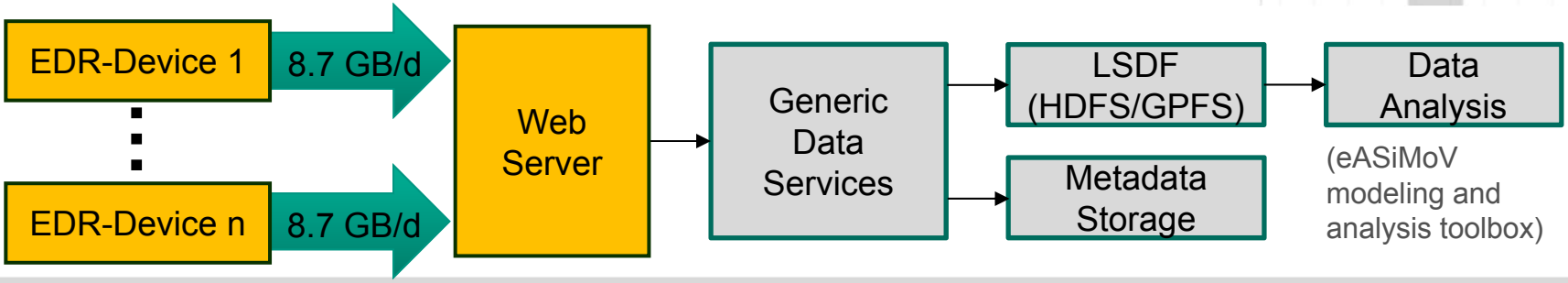






- **DLCL Energy** (KIT, U-Ulm)
  - Secure data management for tech. & eco. data analysis
  - Electrical Data Recorder: 3-phase voltage measurement





LSDMA Experiences

Communities differ in

- Previous knowledge
- Level of specification of the data life cycle
- Tools and services used

Needs driven by

- Increasing amount of data
- Cooperation between groups
- Policies
  - Open access/data
  - Long-term preservation

Within communities

- Focus on data analysis
- High fluctuation of computing experts

Lessons learned

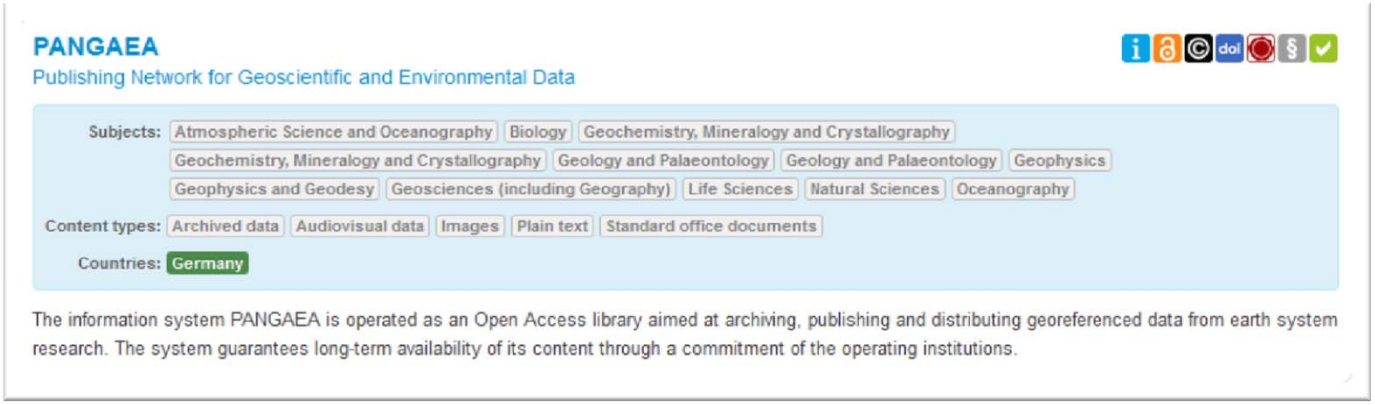
- Interoperable AAI crucial
- Data privacy very challenging, both legally and technically
- Communities need evolution, not revolution
- Needs can be very specific

Re3data.org
© by KIT-BIB, Frank Scholze

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

KIT Karlsruher Institut für Technologie

- Goals
  - Linking existing research data repositories
  - In-depth analysis of quality requirements for research data repositories
  - Define a draft set of criteria for their quality assurance
- Currently 634 research data repositories from around the world covering all academic disciplines are listed
  - 586 of these are described using the re3data.org schema, http://doi.org/10.2312/re3.005



GFZ Helmholtz-Zentrum POTSDAM

Humboldt-Universität zu Berlin

KIT Karlsruhe Institute of Technology

FUNDING DFG

NETWORK DINI (Bibliotheken, Wissenschaft, Rechenzentren, Medienzentren)

Smart Data innovation Lab: A Data Hub for Industry

Smart Data Innovation Lab

Karlsruher Institut für Technologie

Governance

Data Innovation Communities

Industry 4.0

Energy

Smart Cities

Medicine

Working Group **Data Curation**

Working Group **Legal Affairs**

Cross Topic Communities

Working Group
Facility Operation and Tools